An event-driven probabilistic model of sound source localization using cochlea spikes

Anumula, Jithendar ; Ceolini, Enea ; He, Zhe ; Huber, Adrian ; Liu, Shih-Chii

Abstract: This work presents a probabilistic model that estimates the location of sound sources using the output spikes of a silicon cochlea such as the Dynamic Audio Sensor. Unlike previous work which estimated the source locations directly from the interaural time differences (ITDs) extracted from the timing of the cochlea spikes, the spikes are used instead to support a distribution model of the ITDs representing possible locations of sound sources. Results on noisy single speaker recordings show average accuracies of approximately 80% on detecting the correct source locations and an estimation lag of <;100ms.

DOI: https://doi.org/10.1109/ISCAS.2018.8351856

Accepted for publication at the
2018 IEEE International Symposium on Circuits and Systems

An event-driven probabilistic model of sound source localization using cochlea spikes

Jithendar Anumula, Enea Ceolini, Zhe He, Adrian Huber, and Shih-Chii Liu
Institute of Neuroinformatics, University of Zurich and ETH Zurich
Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

***Abstract*—This work presents a probabilistic model that estimates the location of sound sources using the output spikes of a silicon cochlea such as the Dynamic Audio Sensor. Unlike previous work which estimated the source locations directly from the interaural time differences (ITDs) extracted from the timing of the cochlea spikes, the spikes are used instead to support a distribution model of the ITDs representing possible locations of sound sources. Results on noisy single speaker recordings show average accuracies of approximately 80% on detecting the correct source locations and an estimation lag of <100ms.**

***Keywords*—source localization, Bayes filter, silicon cochlea spikes, event-driven auditory processing, probabilistic model**

I. Introduction

With increasing maturity of event-based sensor designs, much work has recently gone into both algorithms and networks that extract input stimulus information from output asynchronous events of event-based sensors such as the retina and cochlea sensors. This development is particularly focused on vision algorithms using the retina events of the Dynamic Vision Sensor (DVS) [1, 2, 3, 4]. Algorithms using cochlea spikes from the Dynamic Audio Sensor are rarer but examples include the use of these spikes in classification tasks such as speaker identification [5] and reconstruction [6].

Some algorithms use the DVS events to drive probabilistic models therefore reducing the impact of the uncertainty of the timing information due to the inherent noise in the spikes. These models have been used, for example, in tasks such as optical flow [7] and image reconstruction [8, 9, 10]. Such models have not been applied as yet to the cochlea spikes.

In this work, we show the application of a probabilistic method for estimating the location of a sound source based on the interaural time difference (ITD) information extracted from the cochlea spikes. Instead of estimating the location based on a histogram of ITDs from the timing of cochlea spikes in response to sounds [11, 12], the extracted ITDs are used to support a probabilistic model of possible source locations. The method was developed for a setup with multiple loudspeakers and was validated using a set of cochlea spike recordings. The work is organized as follows: Section II describes the methods and the recording conditions. Section III describes the experimental results of using the probabilistic model, Section IV describes single speaker separation experiments using the spikes supporting a particular speaker location and Section V presents the discussion.

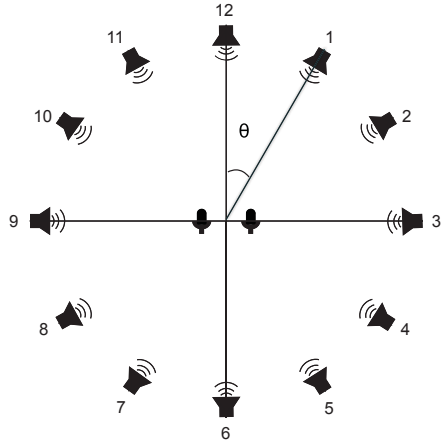

Fig. 1. Measurement setup with twelve speakers used for the audio recordings.

II. Methods

A. Recordings

The silicon binaural cochlea system [12] was used in the recordings for this probabilistic model-based work. This cochlea has two separate 64-stage cascaded filter banks driven by two microphones. Each microphone output goes to a cascaded filter bank modeling the basilar membrane, inner hair cells, and spiral ganglion cells. The circuit details are further described in [12, 13]. The frequency selectivity of the 64 channels range from around 100 Hz to 10kHz. Each channel has 4 neurons.

The recordings were carried out at the University of Zurich, Hospital in a room with 12 loudspeakers arranged in a circle of diameter 3m. Further information about the recording room can be found in the literature [14]. The angular spacing between neighboring speakers is 30 deg. The speakers are numbered from 1 to 12 in a clockwise direction as shown in Fig. 1. The cochlea board was placed in the center of the circle with loudspeaker number 12 in front of the microphones.

TABLE I. Mapping between speaker number and angle

Speaker	12	1	2	3	4	5
Angle	0°	30°	60°	90°	120°	150°
Speaker	6	7	8	9	10	11
Angle	180°	-150°	-120°	-90°	-60°	-30°

The mapping between the speaker number and the angle in the coordinate space of Fig. 1 is shown in Table I. Note

that the ITDs alone are not sufficient to distinguish speakers which are positioned symmetrically about the line through the two microphones, and hence we can only predict seven unique positions in our setup, from -90° to 90°.

The recorded datasets can be divided into three categories: a single speaker dataset, a conversation dataset, and a concurrent dataset. Each categorical dataset contains recordings from both male and female speakers. The single speaker recordings were done using audiobooks. In particular, we collected 5-second excerpts from 8 different audiobooks, 4 narrated by male speakers and 4 narrated by female speakers. From this dataset, the recordings from 3 male speakers and 3 female speakers were used towards training the model which will be described in Section II-C and the remaining two speaker recordings were used towards testing the model. The conversational dataset was created using 15-second excerpts from conversations in the SwitchBoard dataset. In particular, we used sample *sw3762* for Male/Male, sample *sw2017* for Male/Female and sample *sw2014* for Female/Female. Finally, for the concurrent dataset, we superimposed combinations of samples of 10 second each from the single speaker dataset.

a) Single Speaker: A single recording consists of a concatenated file of recorded spikes in response to a 5-second speech played 12 times; each time from one out of 12 different positions, 1 to 12 (see Fig. 1). All the recordings were done once without noise and once with babble noise played from the remaining 11 loudspeakers leading to an SNR of 0 dB. Separate recordings were done for both a female voice and a male voice.

b) Conversation: A single recording consists of a concatenated file of recorded spikes in response to a 15-second conversation played 9 times; each time from a different combination of loudspeakers (l_i^1, l_i^2) where l_i^m is the index of the loudspeaker from where speaker $m \in \{1, 2\}$ was played and $i \in \{1, \ldots, 9\}$ is the combination index. The sequence of positions is the set {(1,11), (2,10), (3,9), (1,12), (3,11), (1,9), (12,11), (2,3), (7,5)}. This sequence was repeated for the combination of Male/Male, Male/Female and Female/Female speakers.

c) Concurrent: These dataset recordings were carried out with the same paradigm described for the conversation recordings.

B. ITD algorithm

To localize a sound source, the recordings from the two ears of the binaural cochlea system were used to estimate the Interaural Time Difference (ITD) values. The ITD value encodes the difference in the arrival time of sound between the two ears. Depending on the position of the sound source, the sound wave takes a different time to reach the two ears. By estimating this ITD from the timing of the spikes, we can infer the position of the source.

While it is possible to accurately determine the ITD from standard microphone recordings by computing the optimal time-shift required to align the waveforms recorded at the two ears, such an estimation technique is not possible with the event-based sensors. The following method was used to estimate the ITD from the silicon cochlea spikes.

For an event e_k at time t_k, frequency channel c_k and at the ear r_k, represented as $e_k = [t_k, c_k, r_k]$, the real-time ITD is estimated by computing the time difference between the current event and the closest event from the same frequency channel but at the other ear. The method is mathematically described as follows:

1) A set of events N_k is built as $N_k = \{t_i \mid |t_i - t_k| \leq T, r_i \neq r_k, c_i = c_k\}$, where T is the maximum ITD considered.
2) Then, a set M_k is constructed from the elements of N_k with the lowest absolute value, i.e. $M_k = \{t \mid t \in N_k, |t| \leq |x| \ \forall x \in N_k\}$.
3) If the set M_k is not empty, the ITD estimate for the current event is then chosen randomly from the set. In case M_k is empty, the event e_k is not used in the estimation of ITD.

In this work, we used a value of $T = 800\mu s$ and only the events from channels 10 to 20 were considered towards the ITD estimation, because these channels gave ITD distributions that were closer to a uni-modal distribution.

C. Probabilistic model

The probabilistic model used for the model-based localization is essentially a Hidden Markov model (HMM), as shown in Fig. 2. The states s_{k-1}, s_k, s_{k+1} refer to the hidden state sequence and o_{k-1}, o_k, o_{k+1} denote the corresponding observations. The arrows in the figure denote the conditional dependency. As shown in the figure, o_k only depends on s_k and s_k only depends on s_{k-1}, which is exactly the Markov assumption. Intuitively, this model illustrates the observation behaviors in a Markov system, where the true states are not directly accessible.

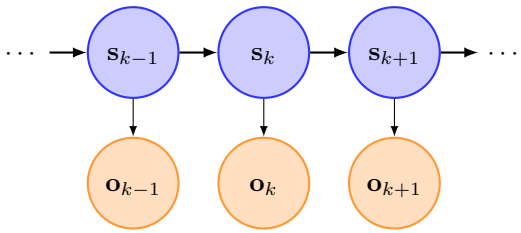

Fig. 2. Hidden Markov Model

$$p(s_k|s_{1:k}) = p(s_k|s_{k-1}) \tag{1}$$

$$p(o_k|s_{1:k}) = p(o_k|s_k) \tag{2}$$

$$p(s_{1:k}, o_{1:k}) = p(s_0) \prod_{i=1}^{k} p(o_i|s_i) p(s_i|s_{i-1}) \tag{3}$$

In our sound localization model, the hidden states describe the true positions of the audio source, and are denoted as s_k, $k \in 1, 2, \ldots$. The observations are the ITD estimates from the silicon cochlea spikes, and are denoted as o_k, $k \in 1, 2, \ldots$.

Integrated with this model, the sound localization problem can be formulated as the determination of the most likely state s_k with the maximal $p(s_k|o_{1:k})$, that is, the posterior distribution of the s_k given k measurements. Based on the

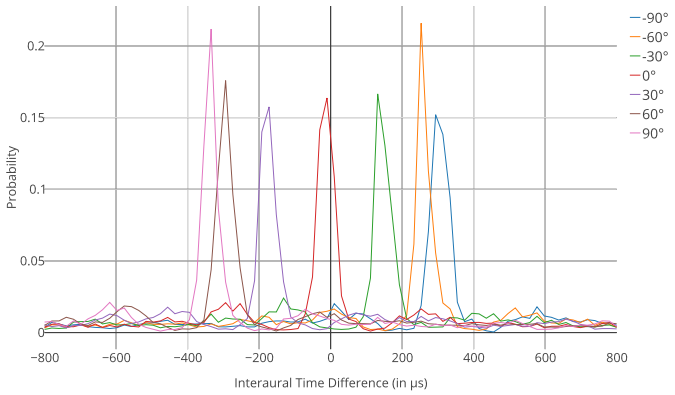

Fig. 3. The extracted ITD distributions from the single male recordings for each of the 7 positions considered in the analysis.

Markov assumption described by Eqs. (1) - (3), we can recursively estimate $p(s_k|o_{1:k})$ through

$$p(s_k|o_{1:k}) \propto p(o_k|s_k)p(s_k|o_{1:k-1}) \tag{4}$$

$$p(s_k|o_{1:k-1}) = \int p(s_k|s_{k-1})p(s_{k-1}|o_{1:k-1})ds_{k-1} \tag{5}$$

Intuitively, Eq. (4) demonstrates that the conditional distribution of s_k given $o_{1:k}$ depends on the likelihood $p(o_k|s_k)$ and the estimation $p(s_k|o_{1:k-1})$ which is based on history $1:k-1$. Eq. (4) gives the update step, where the likelihood $p(o_k|s_k)$ corrects the estimation from the past while Eq. (5) gives the prediction step. The model takes every possible path from $k-1$ to k and estimates the most probable position on account of the last $1:k-1$ ITDs. This algorithm is also known as a Bayes Filter or Recursive Bayes Estimation. It recursively predicts and updates the estimation of the true states. Once the posterior is estimated, the current position estimate is given by the maximum a posteriori (MAP) criterion.

III. Experimental Results

A. Probabilistic approach

The probabilistic approach described in Section II-C needs the likelihood term $p(o_k|s_k)$ and the transition term $p(s_k|s_{k-1})$ so that the update step and the prediction step can be calculated respectively. The likelihood term can be obtained from a set of recordings in a training set with the ground truth, while the transition term $p(s_k|s_{k-1})$ is modelled as a Gaussian centered at s_{k-1}.

The likelihood term is estimated by computing the distribution of ITD estimates from the training set recordings at each of the possible true positions. The distributions of the ITDs can be seen in Fig. 3. The distributions have significant tails and thus could not be properly modelled with Gaussian functions. Instead we discretized the ITD histograms by using time bins of length 20μs.

The term $p(s_k|s_{k-1})$ is the transition probability from one location to the other location, that can be modelled as a Gaussian centered at the angle of the current location with a small variance to account for the noise in the prediction. The variance value was chosen through a hyper-parameter search in this work.

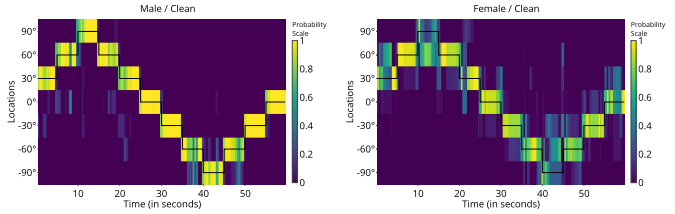

Fig. 4. Probabilities for different locations estimated from the probabilistic model on the (LEFT) single male recording and (RIGHT) single female recording. The ground truth is indicated by horizontal black lines.

The results in Fig. 4 show that the estimated positions are less noisy for the male recordings in comparison to the female recordings. In the results from the male recordings, the estimations of $-60°$ and $-90°$ and those of $60°$ and $90°$ are more likely to be confused. The reason can be seen from the shape of the ITD distributions, where the curves look similar for $-60°$ and $-90°$, and for $60°$ and $90°$. By contrast, the distributions for $0°$, $30°$ and $-30°$ are more separable.

TABLE II. Accuracy results on single speakers in clean and noisy conditions

Speaker Noise	Male Clean	Female Clean	Male Babble	Female Babble
Accuracy per event	93.8%	81.7%	87.8%	76%
Accuracy per 5ms bins	92.3%	84.5%	85.7%	75.9%
Average prediction lag	7ms	12ms	64 ms	93ms

The accuracy results for predicting the correct position of the single speaker are shown in Table II. The predictions by the model lag the ground truth by just about 10ms in the clean datasets and about 80ms in the noisy datasets. Next, the algorithm was used to estimate the locations of two speakers from the conversation dataset and the results are shown in Fig. 5. With two different speakers, each located at different locations, we can observe that the predictions' jumps correlate with the ground truth positions. The estimations on the single speaker datasets were better than the estimations on the conversation dataset because of the sudden jumps to locations not adjacent to each other.

IV. Spike Separation Results

To show that the localization algorithm works also for multi-speaker scenarios, we introduce a spike separation method that shows that spikes assigned to a certain location have indeed been produced by the speaker placed in that location. We evaluate this method with the concurrent dataset described in Section II-A.

For this separation task, we label every spike e_k with the output of the MAP criterion on $p(s_k|o_{1:k})$, thus for the sequence of events ending with event e_k. This assignment allows us to have several spike trains $c_i(t)$ produced by the sources placed at the different positions i.

To validate the separation algorithm, we show that spikes assigned to a certain location have indeed been produced by the speaker placed in that location. This validation is only possible because of the availability of the waveform of the

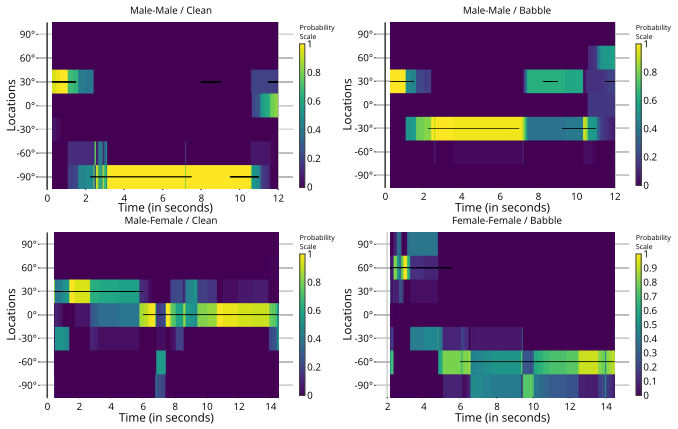

Fig. 5. Estimation of speaker location using the probabilistic model on the conversation dataset. Shown here are the (TOP LEFT) clean male-male, (TOP RIGHT) babble male-male, (BOTTOM LEFT) clean male-female and (BOTTOM RIGHT) babble female-female conversations. The ground truth is indicated by horizontal black lines.

single sources before the mixing. The validation proceeds as follows.

First, we extract the envelope of the waveform by taking the absolute value of its Hilbert transform. Then, we compute a binary vector from the envelope, corresponding to the points of the envelope that match or exceed 10% of the maximal envelope amplitude. This thresholding is done to filter out the low-power segments of the signal that might have not elicited a spike in the cochlea. We then low pass this binary vector by applying a moving window of 20ms and by calculating the number of events, i.e. the number of ones in the binary vector, present in that window to obtain $x_{lp}(t)$. We then use a window of 20ms to calculate a spike count $c_i(t)$ for each trace of spikes assigned to each position l_i for $i \in [0-6]$.

We then compute the Pearson correlation between $x_{lp}(t)$ and each of the $c_i(t)$. The estimated position $\hat{l}_i$ corresponds to the index i for which $\rho(x_{lp}(t), c_i(t))$ is highest following:

$$\hat{l}_i = \underset{i}{\operatorname{argmax}}\, \rho(x_{lp}(t), c_i(t)) \tag{6}$$

We evaluate the spike separation on 3 sets of clean mixtures, namely Male/Male, Male/Female and Female/Female with combinations of positions as described in Section II. For each mixture set, we have 18 classifications, one per speaker in every combination of positions. The results shown in Table III are similar to the results obtained by the localization algorithm when evaluated for spike classification in a scenario with single speaker, proving that the probabilistic localization model also works well in a multi-speaker scenario.

TABLE III. CLASSIFICATION ACCURACY OF SPIKE SEPARATION

Mixture	Male/Male	Male/Female	Female/Female
Accuracy	100%	88.8%	83.3%

An example of spike separation is shown in Fig. 6. The first speaker is shown in the top panel while the second speaker is shown in the bottom one. The spike count for the first speaker is very well correlated ($\rho = 0.5488$) with the envelope of the ground truth, while for the second speaker we see a lower correlation $\rho = 0.3667$. The reason for this is the

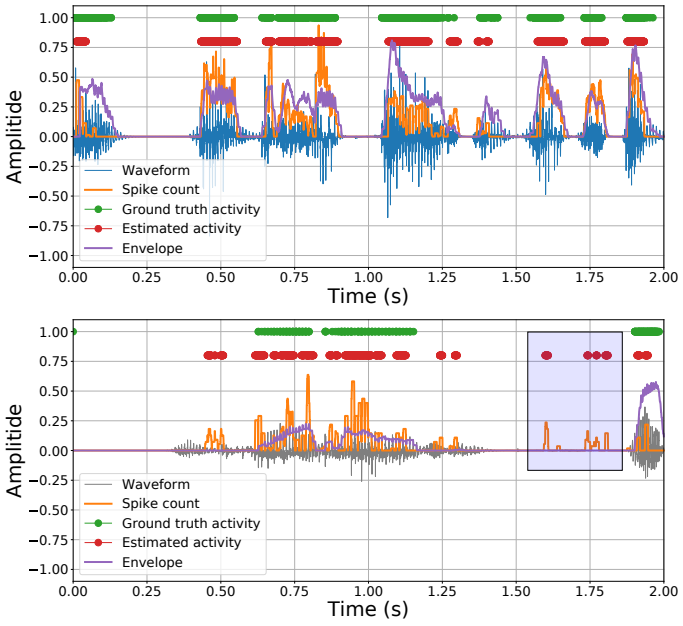

Fig. 6. Example of spike separation in a Male/Male mixture. Curves for the first speaker are shown in the top panel while those for the second speaker are shown in the bottom panel. The waveform (in blue (top) and grey (bottom)) is first smoothed and then thresholded at 10% of its amplitude, to obtain an envelope (violet) and an estimate of the periods of activity (red dots). The spike count extracted from the spikes assigned to the location of the speaker (orange) correlates with a Pearson correlation $\rho = 0.5488$ (top) and $\rho = 0.3667$ (bottom). The light blue zone in the bottom panel shows how certain spikes are wrongly assigned to the second speaker during periods of activity of the first speaker.

presence of some spurious spikes, highlighted by the blue box in Fig. 6. Moreover, as we can see from the colored dots, the spike separation can also be used as a speaker activity detection. While the green dots show the ground truth activity or presence of the speaker, the red dots show the activity estimated by the spike counts.

V. DISCUSSION

This paper presents a probabilistic source localization model applied to silicon cochlea spike recordings. The results on single speaker datasets show that this model predicts the speaker location with an average accuracy of about 80% over a 5ms period in both clean and noisy conditions. In the clean condition, the estimation lag is in the order of 10ms. The localization results for multi-speaker scenarios show an average accuracy of 90% for estimations over a 10s period. This method can be cheaply implemented on mobile platforms that can react to the location of a sound source [15, 16]. The results in this work are currently limited to stationary discrete positions. Future extensions include investigations into Kalman filtering methods for moving sources.

ACKNOWLEDGMENT

The authors would like to thank Prof. Norbert Dillier for the use of the room in which the recordings were carried out, and the assistance of Andrea Kegel with the recording setup. This work was partially supported by the European Union's Horizon 2020 research and innovation program under grant agreement No 644732.

REFERENCES

[1] B. Rueckauer and T. Delbruck, "Evaluation of event-based algorithms for optical flow with ground-truth from inertial measurement sensor," *Frontiers in Neuroscience*, vol. 10, p. 176, 2016.

[2] Z. Ni, S.-H. Ieng, C. Posch, S. Regnier, and R. Benosman, "Visual tracking using neuromorphic asynchronous event-based cameras," *Neural Computation*, vol. 27, no. 4, pp. 925–953, 2015.

[3] E. Mueggler, B. Huber, and D. Scaramuzza, "Event-based, 6-dof pose tracking for high-speed maneuvers," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2014)*, 2014, pp. 2761–2768.

[4] D. Weikersdorfer, R. Hoffmann, and J. Conradt, "Simultaneous localization and mapping for event-based vision systems," in *International Conference on Computer Vision Systems*, 2013, pp. 133–142.

[5] C. H. Li, T. Delbruck, and S. C. Liu, "Real-time speaker identification using the AEREAR2 event-based silicon cochlea," in *2012 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2012, pp. 1159–1162.

[6] M. Yang, C.-H. Chien, T. Delbruck, and S.-C. Liu, "A 0.5V 55μW 64$\times$2 channel binaural silicon cochlea for event-driven stereo-audio sensing," *IEEE Journal of Solid-State Circuits*, vol. 51, no. 11, pp. 2554–2569, 2016.

[7] X. Clady, C. Clercq, S.-H. Ieng, F. Houseini, M. Randazzo, L. Natale, C. Bartolozzi, and R. Benosman, "Asynchronous visual event-based time-to-contact," *Neuromorphic Engineering Systems and Applications*, vol. 51, 2015.

[8] H. Rebecq, T. Horstschaefer, G. Gallego, and D. Scaramuzza, "EVO: A geometric approach to event-based 6-DOF parallel tracking and mapping in real time," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 593–600, April 2017.

[9] M. Cook, L. Gugelmann, F. Jug, C. Krautz, and A. Steger, "Interacting maps for fast visual interpretation," in *2011 International Joint Conference on Neural Networks*, 2011, pp. 770–776.

[10] H. Kim, A. Handa, R. Benosman, S.-H. Ieng, and A. J. Davison, "Simultaneous mosaicing and tracking with an event camera," in *BMVC*, 2014.

[11] H. Finger and S.-C. Liu, "Estimating the location of a sound source with a spike-timing localization algorithm," in *2011 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2011, pp. 2461–2464.

[12] V. Chan, S.-C. Liu, and A. van Schaik, "AER EAR: A matched silicon cochlea pair with address event representation interface," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 54, no. 1, pp. 48–59, 2007.

[13] S.-C. Liu, A. van Schaik, B. A. Minch, and T. Delbruck, "Asynchronous binaural spatial audition sensor with 2$\times$64$\times$4 channel output," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 8, no. 4, pp. 453–464, Aug 2014.

[14] M. F. Mueller, A. Kegel, S. M. Schimmel, N. Dillier, and M. Hofbauer, "Localization of virtual sound sources with bilateral hearing aids in realistic acoustical scenes," *The Journal of the Acoustical Society of America*, vol. 131, no. 6, pp. 4732–4742, 2012.

[15] P. Klein, J. Conradt, and S.-C. Liu, "Scene stitching with event-driven sensors on a robot head platform," in *2015 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2015, pp. 2421–2424.

[16] V. Y.-S. Chan, C. T. Jin, and A. van Schaik, "Adaptive sound localization with a silicon cochlea pair," *Frontiers in Neuroscience*, vol. 4, 2010.